ANALYSIS

52(241), January 1, 2016

ISSN 2278–5469
EISSN 2278–5450

Discovery

Unsupervised feature selection using graph theoretic approach

Venuprabha VS[1], Ajithkumar R[2], Karthikeyan P[3], Anitha N[4]

1. Department of Information Technology, Kongu Engineering College, Perundurai, Erode, India; Email-id: venuvsv@gmail.com
2. Department of Information Technology, Kongu Engineering College, Perundurai, Erode, India; Email-id: karthikeyanit025@gmail.com
3. Department of Information Technology, Kongu Engineering College, Perundurai, Erode, India; Email.id: ajithkumarit2013@gmail.com
4. Assistant Professor (Sr.g), Department of Information Technology, Kongu Engineering College, Perundurai, Erode, India; Email-id: anitha@kongu.ac.in

Article History

Received: 05 November 2015
Accepted: 10 December 2015
Published: 1 January 2016

Citation

Venuprabha VS, Ajithkumar R, Karthikeyan P, Anitha N. Unsupervised Feature Selection Using Graph Theoretic Approach. *Discovery*, 2016, 52(241), 48-54

ABSTRACT

Feature selection is an interesting problem in data mining which can avail to reduce computation time, improve prediction performance, and build understandable models. Concretely, feature selection realized in the absence of class labels, named as unsupervised feature selection, is challenging and fascinating. A method called graph-theoretic approach for unsupervised feature selection has been proposed to solve these issues. The proposed method works in three steps. In the first step, the entire feature set is presented in the form of weighted graph. In the second step, the features are divided into several clusters utilizing a community detection algorithm. All clusters resulted in this approach are relatively independent of each other. So this may useful for most effective results and finally in the third step, a novel iterative search strategy predicated on node centrality is developed to select the final subset of features. The proposed feature selection method provides two major advantages: first, this- method groups features into different clusters predicated on their similarity attributes, in which the features in the same cluster are highly correlated to each other, and to obtain the reduced redundancy set, the reduced subset of features is selected from different clusters. In the second step, the node centrality measure and term variance are employed to identify the most representative and informative feature subset; thus, the optimal size of the subset of features can be automatically determined. This method improves the classification accuracy, reduces the computation time and storage space of the classifier.

Keywords - Feature Selection, Community detection, Laplacian Centrality, Term variance, Influential feature

1. INTRODUCTION

Nowadays datasets have grown hugely and include large numbers of features. Accordingly machine learning methods often deal with samples consisting of thousands of features. A high-dimensional data with irrelevant and redundant features may have harmful consequences in terms of performance, computational cost and storage cost for a given learning task. Feature Selection method reduces the number of original features by selecting a subset of features with sufficient information for classification. The feature selection methods can be classified into four categories including filter, wrapper, embedded, and hybrid models. In the filter-based methods each feature is ranked without consideration of any learning algorithms. The filter model can be broadly classified into univariate and multivariate approaches. In the univariate filter approach each feature is considered separately, thereby ignoring feature dependencies which can effectively identify relevant features independently of any learning algorithms, but they are unable of removing redundant features. The multivariate approaches which considering dependencies between features can handle both irrelevant and redundant features and improves the accuracy of the learning model. The wrapper-based methods apply a learning algorithm can effectively handle both irrelevant and redundant features but it requires high computational cost. In the embedded model the feature selection procedure is considered as a part of the model building process which can handle both irrelevant and redundant features whereas training learning algorithms with large numbers of features will be time-consuming. The goal of the hybrid-based methods is to use computational efficiency of the filter model and proper performance of the wrapper model whereas it may suffer in terms of accuracy, because the filter and wrapper models are considered as two separate steps. The feature selection methods can be classified as supervised feature selection and unsupervised feature selection [1]. Supervised feature selection uses class labels to guide the search process for relevant information and unsupervised feature selection doesn't use any class labels.

Consequently, the feature selection for unsupervised data is a difficult problem and is getting more attention. A feature selection method may be evaluated according to efficiency and effectiveness points of view. The efficiency concerns the time required to find a subset of features whereas the effectiveness is related to the quality of the subset of features. In other words, the unsupervised filter-based methods are typically faster, while the unsupervised wrapper methods usually consider the quality of selected features. To achieve a trade-off between these two issues, a novel unsupervised feature selection method [13] by integrating the concept of graph clustering with the node centrality measures is proposed. To the best of knowledge, this is the first paper to apply the node centrality measure to the problem of feature selection. The proposed method, called the Graph Clustering with Node Centrality for unsupervised feature selection, in short GCNC does not need any learning algorithms or class labels to select feature subsets; therefore, it can be classified as an unsupervised filter-based approach and will be computationally efficient for high-dimensional datasets.

2. RELATED WORKS

Cheng et al [4] presented a spectral semi-supervised feature selection criterion called s-Laplacian score. Based on this criterion they proposed a Graph-based Semi-supervised Feature Selection method (GSFS). Here, the spectral graph theory and conditional mutual information are utilized to select relevant features and also remove redundant features. Hsu et al [5] presented a hybrid approach which is a combination of the filter and wrapper models and it attempts to take advantage of both approaches. This approach mainly focuses on combining the filter-based and the wrapper-based methods to achieve the best possible performance with a particular learning algorithm and time complexity similar to the filter-based methods. Zhang et al [9] developed a hyper graph based information theoretic method for feature selection. Here, feature space was represented by a hyper graph where each node denoted feature, and each edge had a weight corresponding to the multidimensional interaction information between the features connected by the edge, and then, a specific hyper graph clustering algorithm was applied to the hyper graph in order to locate the most informative feature subset. Song et al [11] a Fast clustering- based feature election algorithm for high dimensional data is proposed. This method works in four different steps including: (1) removing irrelevant features, (2) constructing minimum spanning tree from relative ones, (3) partitioning the spanning tree, and (4) selecting representative features. Bandyopadhyay et al [12] an unsupervised feature selection method has been developed by integrating the concept of densest sub-graph finding with feature clustering. This method works in two steps: in the first step, the densest sub-graph is obtained so that the features are maximally non-redundant among each other and in the second step; a specific feature clustering algorithm is performed around then on redundant features in order to produce a reduced feature set.

3. A GRAPH THEORETIC APPROACH

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. The proposed method works in three steps:

Step 1: The problem space is represented as a graph by considering the entire feature set as the vertex set and having feature similarity as the corresponding edge weight.
Step 2: Features are divided into several clusters by employing a community detection method.
Step 3: A novel iterative search process is used by utilizing the node centrality and term variance [15]. The proposed method can deal with both irrelevant and redundant features.

The Fig 3.1 represents the flow diagram of the proposed system. It shows how the high dimensional dataset is reduced into reduced subset of features. This is effectively done by integrating the concept of graph clustering with node centrality measures.

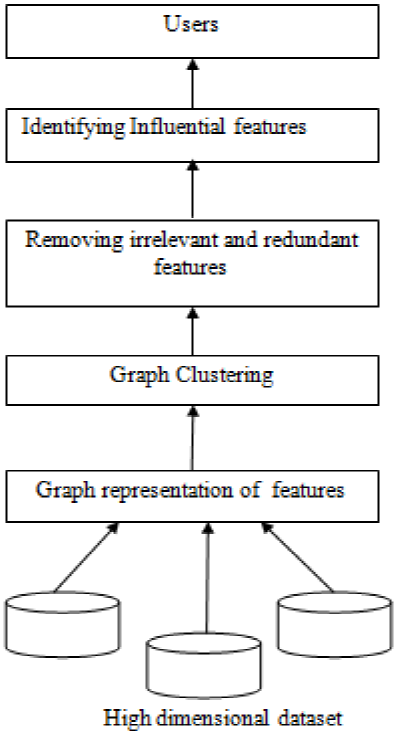
Users
Identifying Influential features
Removing irrelevant and redundant features
Graph Clustering
Graph representation of features
High dimensional dataset

Figure 3.1 Flow diagram of graph theoretic approach

3.1. Graph representation of features

A preliminary step for all graph-based methods is to represent training data with an undirected graph. For this purpose, the feature set is mapped into its equivalent graph G={F, E , W_F}, where F = {F_1, F_2, F_3,.......,F_n} is a set of original features, E={ (Fi , Fj) : Fi , Fj ϵ F} denotes the edges of the graph and w_{ij} indicates the similarity between two features F_i and F_j connected by the edge (F_i , F_j). Pearson product–moment correlation coefficient [7] is used to measure the similarity between different features of a given training set. The correlation between two features F_i and F_j is defined as follows:

$$w_{ij} = \left| \frac{\sum_p (x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{\sum_p (x_i - \bar{x}_i)^2}\ \sqrt{\sum_p (x_j - \bar{x}_j)^2}} \right| \qquad (1)$$

In Equation (1), x_i and x_j denote the vectors of features F_i and F_j, respectively. Variables x_i and x_j represent the mean values of vectors x_i and x_j averaged over p samples. It is clear that the similarity value between a pair of features which are completely similar will be equal to 1, and for completely dissimilar features this value will be equal to 0.

3.2 Feature Clustering

Feature clustering is an efficient approach for dimensionality reduction [6]. A community detection method called the Louvain community detection algorithm is used to identify the feature clusters. This algorithm detects communities in the graph by maximizing a specific modularity function. It should be noted that before using the clustering method, the edges with associated weights lower than a threshold θ will be removed in order to improve performance of the Louvain community detection algorithm. The θ parameter can be set to any value in the range [0 1]; thus when its value is small (large), more (fewer) edges will be considered in the graph clustering algorithm and the number of obtained clusters will be low (high).

3.3. Community detection

The goal of community detection is to cluster the similar vertices into one community separate from others. The Louvain community detection algorithm [2] is an algorithm for performing community detection (i.e. clustering) in networks by maximizing a modularity function. This method is a simple, efficient and easy-to-implement one for identifying communities in large networks in short computing times. It is based on two simple steps: In the first step each node is assigned to a community chosen in order to maximize the network modularity Q. The gain derived from moving a node i into a community C can simply be calculated as follows:

$$\Delta Q = \frac{\Sigma_c + k_i^c}{2m} - \left(\frac{\Sigma\hat{c} + k_i}{2m}\right)^2 - \left[\frac{\Sigma_c}{2m} - \left(\frac{\Sigma\hat{c}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)\right] \qquad (2)$$

In Equation (2), Σ_C is the sum of the edge weights inside C, $\Sigma\hat{c}$ is the sum of the edge weights incident to nodes in C, k_i is the sum of the edge weights incident to node i, k_i^C is the sum of the edge weights from i to nodes in C, and m is the sum of all the edge weights in the network.

The second step simply makes a new network consisting of nodes of the communities previously found. Then the process iterates until a significant improvement of the network modularity is obtained.

3.4. Representative features selection

Here, a specific iterative search process removes features with influential values less than a predefined threshold δ. After removing these features, the clustered graph is reconstructed. This process is repeated until there is no valuable feature with an influential value less than δ or the number of features in a specific cluster is greater than a predefined threshold ω. To measure the influential value of each feature, integration of node centrality and term variance is performed. The influential value [13] of feature F_i, denoted as Inf (F_i), is defined as follows:

$$\mathrm{Inf}(F_i) = \mathrm{LC}(F_i) \times \mathrm{TV}(S, F_i) \qquad (3)$$

In Equation (3), LC(Fi) denotes the Laplacian centrality of feature Fi and TV(S,Fi) indicates the normalized term variance of feature Fi that is defined as follows:

$$\mathrm{TV}(S, F_i) = \frac{1}{|S|}\sum_{j=1}^{|S|}\left(A_{ij} - \overline{A_i}\right)^2 \qquad (4)$$

In Equation (4), A_{ij} indicates the value of feature Fi for the pattern j, and |S| is the total number of patterns.

3.5. Node centrality

The Laplacian centrality measure [8] is applied to measure the node centrality. For weighted graph G, W and X matrices are defined as follows:

$$W(G) = \begin{pmatrix} 0 & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & 0 & \cdots & w_{2,n} \\ . & . & . & . \\ w_{n,1} & w_{n,2} & \cdots & 0 \end{pmatrix}$$

$$X(G) = \begin{pmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ . & . & . & . \\ 0 & 0 & \ldots & X_n \end{pmatrix}$$

where,

$$X_i = \sum_{j=1}^{n} W_{ij} = \sum_{u \in N(v_i)} W_{v_i,u} \qquad (5)$$

The Laplacian Energy of G is defined as the following invariant:

$$E_L(G) = \sum_{i=1}^{n} X_i + 2 \sum_{i<j} W_{ij}^2 \qquad (6)$$

Finally, the Laplacian Centrality $C_L(v_i, G)$ of vertex vi is defined as follow:

$$C_L(v_i, G) = \frac{(\Delta E)_i}{E_L(G)} = \frac{E_L(G) - E_L(G_i)}{E_L(G)} \qquad (7)$$

where G_i is the graph obtained by deleting vi from G. The complexity of computing Laplacian centrality for network G with n vertices and maximum degree Δ would be no more than O (n.Δ^2)

4. EXPERIMENTAL RESULTS

4.1. Used Datasets

The basic characteristics of the datasets are,

Table 4.1 Dataset Description

Dataset	Features	Classes	Patterns
Wine	13	3	178
Hepatitis	19	2	155
WDBC	30	2	569
Ionosphere	34	2	351
Spambase	57	2	4601
Sonar	60	2	208
Arrhythmia	279	16	452
Colon	2000	2	62

The performance of the proposed system is analyzed in terms of number of features selected and classification accuracy of original feature set and reduced feature set.

Table 4.2 Number of features selected by the system

Dataset	GCNC	All Features
Wine	6	13
Hepatitis	6	19

WDBC	8	30
Ionosphere	16	34
Spambase	26	57
Sonar	24	60
Arrhythmia	18	279
Colon	39	2000

Table 4.3 Classification Accuracy with reduced subset of features

Dataset	Classification Accuracy with reduced subset				
	IB1	Simple Cart	Random Tree	Decision Table	Naive Bayes
Wine	94.42	93.44	95.08	95.06	94.53
Hepatitis	86.22	83.76	85.09	85.65	89.64
WDBC	95.34	92.37	94.29	96.84	94.63
Ionosphere	88.90	87.72	89.91	89.81	88.72
Spambase	88.21	89.05	88.11	89.46	89.27
Sonar	76.33	78.72	78.34	79.87	79.87
Arrhythmia	69.08	69.99	79.54	79.34	81.54
Colon	82.37	81.43	83.46	81.90	89.68

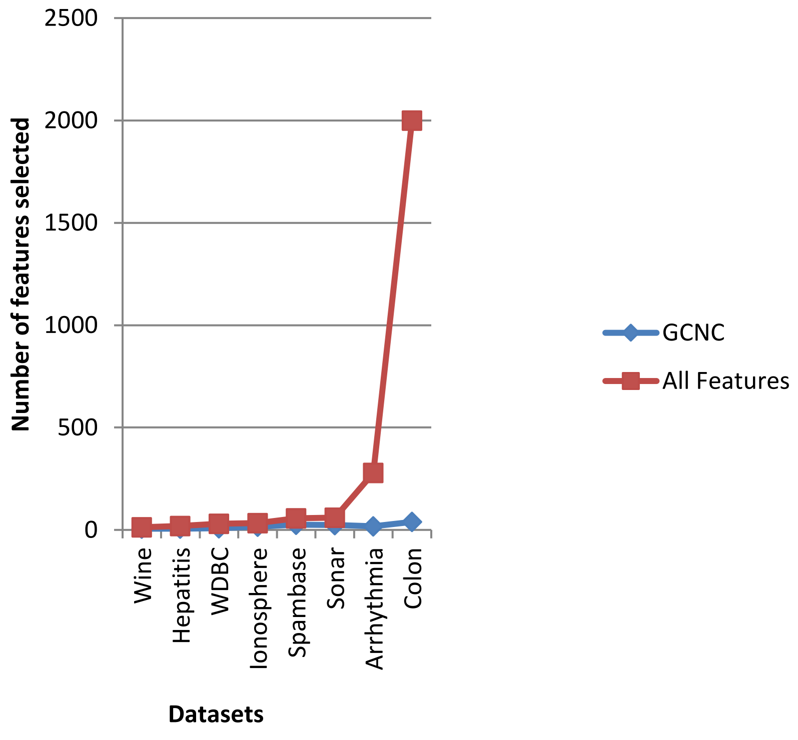

Figure 4.1 Number of features selected from each dataset

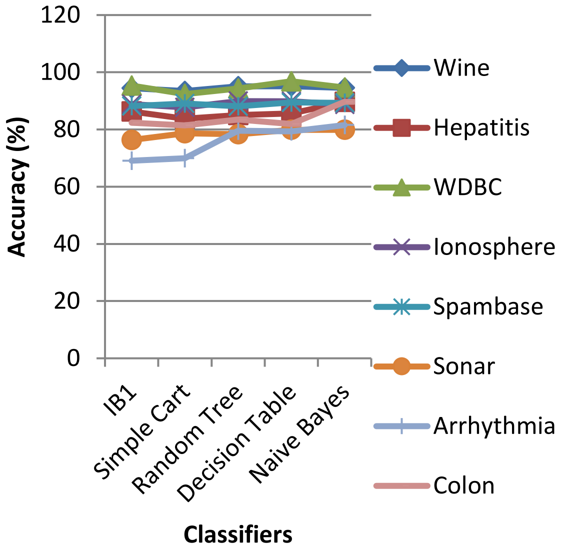

Figure 4.2 Classification Accuracy with reduced subset

5. CONCLUSION

The important problem addressed in the unsupervised feature selection method is less classification accuracy, high computation time, large storage space, computational complexities and inefficiently dealing with irrelevant and redundant features. These issues can be solved by integrating the concept of graph clustering with the node centrality. The proposed method works in three steps: in the first step, the problem space is represented as a graph by considering the entire feature set as the vertex set and having feature similarity as the corresponding edge weight. In the second step, features are divided into several clusters by employing a community detection method. Finally, in the third step, an iterative search process is used by utilizing the node centrality and term variance. The proposed method effectively deals with both irrelevant and redundant features. The proposed method has been analyzed with well known classifiers. The reported results show that in most cases the proposed method obtained the best classification accuracy. Furthermore, the results indicate that the execution time of the proposed method is comparable to those of the other feature selection methods.

REFERENCE

1. He, Cai, Deng and Niyogi1 P (2005), 'Laplacian score for feature selection', Adv. Neural Inf. Process. Syst., Vol.18, pp.507–514.
2. Blondel V, Guillaume J, Lambiotte R and Lefebvre E (2008), 'Fast unfolding of communities in large networks', J. Stat. Mech.: Theory Exp., Vol.10008, pp.1–12.
3. Theodoridis S and Koutroumbas K (2008), Pattern Recognition. Academic Press, Oxford.
4. Cheng, Hongrong, Deng, Wei, Fu, Chong, Wang, Yong and Qin Z (2011), 'Graph based semi-supervised feature selection with application to automatic spam image identification', Comput. Sci. Environ. Eng. Eco Inform., Vol.159, pp.259–264.
5. Hsu, Hui-Huang, Hsieh, Cheng-Wei and Lu (2011), 'Hybrid feature selection by combining filters and wrappers', Expert Syst. Appl., Vol. 38, pp. 8144–8150.
6. Jiang, Jung-Yi, Liou and Ren-Jia (2011), 'A fuzzy self-constructing feature clustering algorithm for text classification', IEEE Trans. Knowl. DataEng., Vol.23, pp.335–349.
7. Monirul Kabir, M D, Shahjahan MD, and Murase K (2011), 'A new local search based hybrid genetic algorithm for feature selection', Neuro computing Vol.74, pp.2914–2928.
8. Qi, Xingqin, Fuller, Eddie, Wu and Zhang CQ (2012), 'Laplacian centrality: a new centrality measure for weighted networks', Inf. Sci., Vol.194, pp.240–253.
9. Zhang Z and Hancock E R (2012), 'Hyper graph based information theoretic feature selection', Pattern Recognit.Lett., Vol.33, pp.1991–1999.
10. Li, Yan, Xiangbin, Zhaia and Fan W (2013), 'C-index: a weighted network node centrality measure for collaboration competence', Informetr., Vol.7, pp.223–239.
11. Song Q, Ni J and Wang G (2013), 'A fast clustering-based feature subset selection algorithm for high-dimensional data', IEEE Trans. Knowl. Data Eng., Vol.25, pp.1–14.
12. Bandyopadhyay S, Bhadra T and Mitra P (2014, 'Integration of dense sub graph finding with feature clustering for unsupervised feature selection' Pattern Recognit. Lett., Vol.40, pp.104–112.
13. Parham Moradi and Mehrdad Rostami (2015), 'A graph theoretic approach for unsupervised feature selection', Eng. Appl. of Artificial Intelligence, Vol.44, pp.33-45.